Journal of the Korea Academia-Industrial cooperation Society
Vol. 16, No. 1 pp. 726-734, 2015

http://dx.doi.org/10.5762/KAIS.2015.16.1.726
ISSN 1975-4701 / eISSN 2288-4688

하둡 기반 빅 데이터 기법을 이용한 웹 서비스 데이터 처리 설계 및 구현

김현주[1*]
[1]단국대학교 대학원 전자전기공학부 컴퓨터응용 전공

Design and Implementation of an Efficient Web Services Data Processing Using Hadoop-Based Big Data Processing Technique.

Hyun-Joo Kim[1*]
[1]**Dept. of Electronics&Electrical Engineering Graduate School Dan-kook University**

요 약 데이터를 구조화하여 사용하는 관계형 데이터베이스가 현재까지 데이터 관리에 가장 많이 사용되고 있다. 그러나 관계형 데이터베이스는 데이터가 증가되면 데이터를 저장하거나 조회할 때 읽기, 쓰기 연산 수행에 제약 조건이 발생되어 서비스가 느려지는 현상이 나타난다. 또 새로운 업무가 추가되면 데이터베이스 내 데이터는 증가되고 결국 이를 해결하기 위해 하드웨어의 병렬 구성, CPU, 메모리, 네트워크 등 추가적인 인프라 구성을 필요로 하게 된다. 본 논문에서는 관계형 데이터베이스의 데이터 증가로 느려지는 웹 정보서비스 개선을 위해 기존 관계형 데이터베이스의 데이터를 하둡 HDFS로 전송하고 이를 일원화하여 데이터를 재구성한 후 사용자에게 하둡 데이터 처리로 대량의 데이터를 빠르고 안전하게 추출하는 모델을 구현한다. 본 시스템 적용을 위해 웹 기반 민원시스템과 비정형 데이터 처리인 이미지 파일 저장에 본 제안시스템을 적용하였다. 적용결과 관계형 데이터베이스 시스템보다 제안시스템 데이터 처리가 0.4초 더 빠른 결과를 얻을 수 있었고 기존 관계형 데이터베이스와 같은 대량의 데이터를 처리를 빅 데이터 기법인 하둡 데이터 처리로도 웹 정보서비스를 지원이 가능하였다. 또한 하둡은 오픈소스로 제공되어 소프트웨어 구매 비용을 줄여주는 장점이 있으며 기존 관계형 데이터베이스의 데이터 증가로 효율적인 대용량 데이터 처리를 요구하는 조직에게 도움을 줄 수 있을 것이다.

Abstract Relational databases used by structuralizing data are the most widely used in data management at present. However, in relational databases, service becomes slower as the amount of data increases because of constraints in the reading and writing operations to save or query data. Furthermore, when a new task is added, the database grows and, consequently, requires additional infrastructure, such as parallel configuration of hardware, CPU, memory, and network, to support smooth operation. In this paper, in order to improve the web information services that are slowing down due to increase of data in the relational databases, we implemented a model to extract a large amount of data quickly and safely for users by processing Hadoop Distributed File System (HDFS) files after sending data to HDFSs and unifying and reconstructing the data. We implemented our model in a Web-based civil affairs system that stores image files, which is irregular data processing. Our proposed system's data processing was found to be 0.4 sec faster than that of a relational database system. Thus, we found that it is possible to support Web information services with a Hadoop-based big data processing technique in order to process a large amount of data, as in conventional relational databases. Furthermore, since Hadoop is open source, our model has the advantage of reducing software costs. The proposed system is expected to be used as a model for Web services that provide fast information processing for organizations that require efficient processing of big data because of the increase in the size of conventional relational databases.

Key Words : Big Data, Hadoop, HDFS, Web Service, RDBM

*Corresponding Author : Hyun-Joo Kim(Dan-kook Univ.)
Tel: +82-10-3590-3731 email: chopinkhj@gmail.com
Received October 28, 2014 Revised (1st December 5, 2014, 2nd December 8, 2014) Accepted January 8, 2015

1. 서론

인터넷 존재하는 데이터가 현재까지 1ZB를 넘어선다고 IDC(International Data Corporation)는 2011년 보고서를 통해 말하고 있다. 최근 빅 데이터(Big Data)는 큰 이슈로 떠올라 그 기술에 관심을 두고 있으며 전 세계적으로 IT분야에 큰 화두가 되고 있다. 우리나라에서도 국가정보화를 데이터 기반으로 변경하는 등 빅 데이터는 차세대 산업 기술로 주목받고 있다[1]. 이처럼 빅 데이터가 기업경쟁력이나 국가경쟁력을 좌우하는 주요 자산으로 여겨지는 이유는 기존의 일부 데이터를 분석하여 얻어지는 결과보다 분석 결과의 정확도가 높으며 이로 인해 기존에 몰랐던 새로운 사실을 발견 할 수 있기 때문이다. 또한 빅 데이터는 향 후 국가 기반 주요 기술로 성장 할 것이라는 것에 대해서는 모두가 의심하지 않고 있다[2-5]. 특히 스마트폰, 태블릿 pc 등 스마트기기와 소셜미디어를 통해 수집되는 데이터 증가로 내・외부 데이터의 서비스 연동과 신속한 정보서비스 지원을 위한 일관된 데이터 관리는 모든 기관이 공통적으로 고민하는 분야이다.

대부분의 기관은 기관 내 업무시스템을 RDBMS(Relational Database Management System)로 데이터를 관리한다. RDBMS는 현재까지 데이터 관리에 가장 많이 활용되는 기술로 엔터프라이즈 컴퓨팅을 사용하는 분야에서는 근 20여년간 관계형 데이터베이스에 데이터를 저장하고 있다. 현재까지 가장 유용하게 사용하는 RDBMS는 저장 할 데이터가 많거나 조회할 데이터가 많아지면 읽기(Read), 쓰기(Write) 연산 수행에 제약 조건이 발생되어 서비스가 느려지는 단점이 있다[2][6-7]. 또 업무가 증가되어 새로운 업무가 추가되면 업무별 RDBMS는 증가되고 이를 지원하기 위한 인프라 자원도 계속적으로 증가하게 된다. 결국 더 많은 CPU, 더 넉넉한 메모리, 고속의 디스크를 탑재한 신규서버 도입하거나 또는 데이터베이스 서버를 추가하여 병렬로 구성하는 등 데이터 관리를 위한 경제적 비용은 계속적으로 증가하게 된다.

본 논문에서는 현재까지 용이하게 사용되어 온 RDBMS 데이터 운영 방식에서 벗어나 효율적 데이터 처리와 경제적으로 유용한 데이터 처리 방식의 연구에 관심을 두었다. 이를 활용하기 위해 하둡(Hadoop)HDFS(Hadoop Distributed File System) 파일 처리를 이용하였다. 각 기관에서 사용하는 각종 RDBMS 데이터를 자동화 스케줄러에 의해 데이터 전처리기로 전송하고 전처리기에 수집 된 데이터는 빅 데이터 처리 기법인 하둡 HDFS로 일원화한다. 이를 맵리듀스를 이용하여 병렬로 재구성하여 대량의 데이터를 빠르고 안전하게 추출하는 데이터 처리 모델을 설계하였다. 하둡을 이용한 데이터 처리는 대용량 파일을 저장할 수 있는 분산 파일시스템으로 제공하여 클러스터로 구성하며 이를 멀티 노드로 부하 분산 처리하므로 시스템의 과부하나 병목현상을 줄여 줄 수 있다[8]. 무엇보다 하둡은 오픈소스로 제공되어 경제적 비용을 줄여주는 큰 장점을 가지고 있다. 본 제안 시스템은 기존 관계형 데이터베이스의 데이터 증가로 효율적인 대용량 데이터 처리를 요구하는 조직에게 신속한 정보서비스 처리를 제공하는 웹 서비스 모델이 될 수 있을 것이다.

2. 빅 데이터

빅 데이터라는 신 개념의 데이터 출현은 스마트기기, 소셜미디어와 더불어 현대 사회의 문화를 바꾸고 경제, 산업계에서는 비즈니스를 변화시키고 있다.

2.1 빅 데이터의 정의

민간 부분에서 빅 데이터 플랫폼 기술 개발에 가장 먼저 투자를 시작한 IBM은 오늘날 매일 2.5 퀸틸리언 바이트(2.5 quintillion bytes = 2.5×10^{18} bytes)의 데이터가 생산된다. 이런 데이터를 “센서 정보, 소셜미디어 사이트 웹문서, 디지털 사진과 동영상, 구매기록, GPS신호” 등 모든 곳에서 생산되는 데이터를 “빅 데이터”로 정의하고 있다[9]. 또한 가트너는 빅 데이터에 대한 정의를 “빅 데이터는 크기가 크고 속도가 빠르며 다양한 정보 자산을 가지고 있다”라고 말하며 맥킨지는 “빅 데이터란 전형적인 데이터베이스로는 다루기 힘든 크기의 데이터 셋으로 빅 데이터를 특정 크기로 지칭 할 수 없다”로 정의하고 있다[10-11].

2.2 빅 데이터의 특징

빅 데이터의 구성은 규모(volume), 형태(variety), 속도(velocity) 3가지 속성으로 구성 되며 이 3가지 속성이 충족 될 때 빅 데이터의 구성이 가능하다. 이 3가지 속성

을 기준으로 빅 데이터의 특징을 구분한다.

첫째, 데이터의 양(Volume)이다. 데이터의 용량은 시간이 흐를수록 증가된다. 빅 데이터의 대용량과 데이터의 지속적인 증가는 데이터 분석 기술의 발전을 요구한다. 둘째, 다양성(Variety)이다. 빅 데이터 수집의 원천은 웹, 소셜미디어 그리고 데이터소스의 로그 및 클릭스트림 등의 정보이다. 이들 정보로 다양한 분야에서 기존 데이터에서 찾을 수 없는 새로운 정보를 얻을 수 있다. 셋째, 속도(velocity)이다. 데이터는 과거와 다르게 기하급수적으로 증가되고 있다. 이들 데이터는 생성과 동시에 즉시 수집된다. 그러므로 수집시간은 급격히 단축된다. 그 외에도 IBM은 정확성을 추가하여 구분하기도 한다.

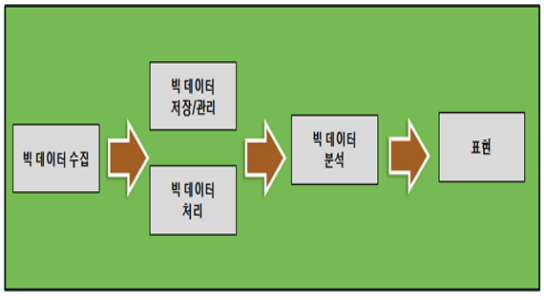
2.3 빅 데이터 컴퓨팅 인프라

빅 데이터 처리에는 고 확장성, 고성능 컴퓨팅 인프라가 없다면 대량의 데이터를 고속으로 처리할 수 없다. 그러므로 대량의 데이터를 빠른 시간에 수집하여 정보의 가치를 얻어 이용하는 기술을 필요로 한다. 또한, 빅 데이터의 활용도에 따라 빅 데이터의 가치를 높일 수 있다. 맥캔지의 2011년 보고서에도 빅 데이터는 공공・행정, 의료・건강, 개인정보, 유통・소매, 제조업 등에서 22.3조 달러의 활용 가치를 예측하고 있다. 빅 데이터의 가치를 높이기 위한 컴퓨팅 인프라 기술로는 분산컴퓨팅, 고성능컴퓨팅, 인-메모리 기술이 있으며, 이를 빅 데이터 컴퓨팅 인프라의 핵심적인 요소기술(Element Technology)이라 한다[2].

2.4 빅 데이터의 활용

빅 데이터의 활용은 다음의 단계를 거친다. 생성→수집→저장→분석→표현의 단계를 거쳐 각 단계별 기술이 적용된다.

먼저 데이터를 생성・수집하고 수집된 데이터는 전처리 과정으로 데이터를 필터링하거나 적절한 형태로 가공을 한다. 가공된 데이터를 체계적으로 저장 관리하여 그 중 유용한 자료는 정보처리 분석과정에서 데이터의 가시화 및 시각화를 통해 활용이 가능하도록 한다[2]. 빅 데이터의 데이터 수집은 데이터 소스로 부터 시작된다. 일반적으로 데이터 수집은 내부정보시스템에 저장된 정형화된 데이터를 말한다. 정형화된 내부 데이터는 업무 수행 과정을 통해 자동으로 수집된다. 그러나 빅 데이터는 내부데이터 외에도 외부에 존재하는 무한한 데이터도 수집하고 수집된 정보를 분석하여 특정 데이터로 변환하는 과정을 거쳐야 한다. 빅 데이터를 저장하는 기술로는 분산 파일시스템(DFS: Distributed File System), NoSQL(Not Only SQL), 메모리 기반 데이터베이스가 있다. 분산파일시스템은 막대한 양의 데이터를 저장 관리하기 위해 물리적으로 서로 다른 컴퓨터에 데이터를 나누어 저장하고 관리하는 파일시스템이다. 빅 데이터의 대용량 데이터 처리를 위해 분산처리 기술인 하둡 HDFS은 빅 데이터 기본 기술로 사용된다. 기존에는 데이터를 처리를 할 때 그 종류와 특성을 미리 정해 놓고 데이터 처리를 했었다. 빅 데이터 처리에서는 데이터의 다양성과 데이터의 크기·용량에 따라 데이터를 처리하는 것이 주요과제이다. 빅 데이터의 데이터 처리 방법은 잘 저장된 데이터를 처리하는 일괄 처리방법과 새로이 생성되어 저장되기 전에 실시간으로 처리하는 방법이 있다. 빅 데이터의 분석은 금융, 공공분야, 범죄검출, 이벤트 기반 마케팅, 소셜미디어 분석, 그 외 다양한 비즈니스 분야에 적용에 된다[12]. 또한, 빅 데이터 분석 방법의 대부분은 기존 통계학과이나 전산학에서 사용되던 데이터마이닝, 기계학습, 자연언어 처리, 패턴 인식 등이 사용된다. 특히 통계처리를 위한 공개 소프트웨어로 R이 최근에 가장 주목 받는 분석 도구이다. R은 통계그래픽 기능이 매우 우수한 오픈소스로 분석된 빅 데이터를 표현하여 데이터 시각화(Data Visualization)에 이용된다. [그림 1]은 빅 데이터 활용을 위해 데이터를 수집하여 최종 활용까지의 진행 절차 그림이다.

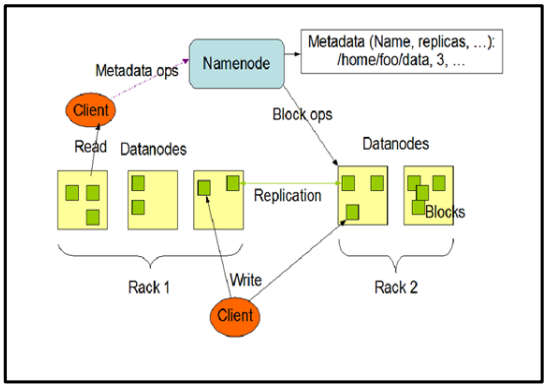


[Fig. 1] Procedure from collection to use of Big data

2.5 하둡(Hadoop)

하둡은 오픈소스로 대규모 데이터의 분산 처리 기술을 지원한다. 특히 대량의 비 구조화 데이터 처리 성능이 뛰어나며 비용이 저렴하며 스케일 아웃 구조로 대용량

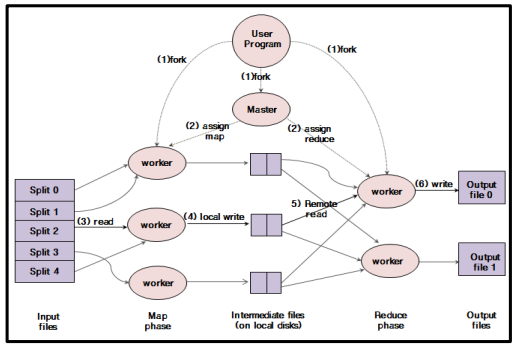
증가에 용이해 주목을 받고 있다. 또한 하둡은 노드의 추가 제거가 용이하고 가용성(Availability)이 높아 일부 장비에 장애가 발생하더라도 전체 시스템에는 영향을 주지 않는다. [그림 2]는 HDFS의 구조로 NameNode와 DataNode의 역할을 도식화한 그림이다. HDFS는 Master인 NameNode와 Slave인 DataNode로 구성된다. NameNode는 파일의 메타(meta) 정보만 관리하고 실제 데이터는 다수의 DataNode에 저장되며 하나의 Secondary NameNode와 연결되어 NameNode의 네임스페이스 정보를 재 저장한다[13-14].



[Fig. 2] Hadoop Distributed File System Structure

2.6 맵리듀스(MapReduce)

맵리듀스는 대용량 데이터 처리, 생성을 위한 프로그래밍 모델이다. 대용량의 클러스터 범용시스템에서 병렬처리되어 자동으로 실행된다. 따라서 사용자가 병렬 및 분산처리에 익숙하지 않아도 대규모 분산시스템을 쉽게 활용할 수 있으며 프로세싱을 작은 단위의 작업으로 세분화하여 클러스터 내 수백 개의 노드에서 병렬로 실행할 수 있다[15-16]. [그림 3]은 맵리듀스의 실행 흐름도이다. 맵리듀스의 실행 과정은 User Program은 시스템명령 fork()를 이용하여 분산 실행된다. 그 중 하나는 Master로 동작하고 나머지는 Map과 Reducedp 동작되는 work를 생성한다. Map에 의해 할당된 worker는 분할된 데이터를 읽어 중간파일 형태 (k2, v2)를 생성한다. 이 과정이 Map과정이며 다시 중간 결과는 local Disks에 저장된다. 이 때 데이터가 Reduce 하는 worker에 의해 다시 취합되어서 결과 파일(output file0, output file2)이 생성된다.



[Fig. 3] Map-Reduce Implementation Process

3. 빅 데이터 프레임워크를 이용한 데이터 처리 모델

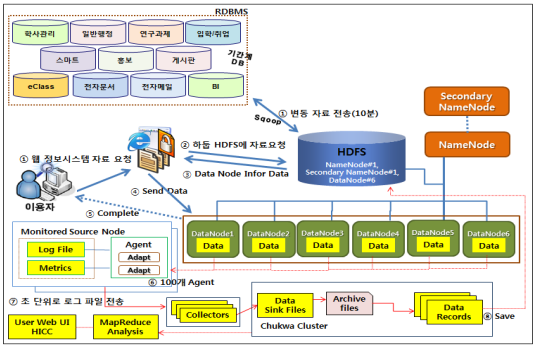
3.1 기존 데이터 처리의 문제점

대부분의 정보시스템은 RDBMS로 데이터를 구성하여 사용하고 있다. 그러나 RDBMS는 일정기간 사용 후 데이터양이 증가되면 읽기(Read), 저장(Write) 연산 수행에 제약 조건이 발생되어 서비스가 느려지는 현상이 발생된다. 이를 해결하고자 CPU, 메모리, 고속의 디스크 탑재, 신규서버 도입 등 서버 인프라에 재투자를 진행하거나 RDBMS 튜닝(Tuning)을 통해 사용자는 해결책을 찾고자 한다. 전자의 경우 계속적인 경제적 비용이 투자되어야 하고 후자의 경우는 기술자의 스킬과 기관의 꾸준한 교육 지원을 통해 얻어지는 결과이기도 하다. 그러나 이 방법은 기하급수적으로 증가되는 데이터 운영 환경 즉, 기존의 RDBMS 운영 환경에서는 적절한 해결책이라 볼 수는 없다는 것이다. 특히 영세한 기관에서 정보시스템에 대한 재투자는 경제적 비용을 가중시켜 큰 부담으로 작용하기 때문이다.

3.2 빅 데이터 프레임워크를 이용한 데이터 처리 모델

본 논문에서 하둡을 이용한 데이터 처리 모델을 제안한다. 본 논문의 데이터 처리는 전통적인 RDBMS 데이터베이스의 업무 효율성 증진과 기존 인프라의 추가 시설 없이 서비스 증가로 인한 데이터 읽기 부하, 쓰기 부하를 개선해 보고자 시도된 데이터 서비스 모델이다. 본 모델은 RDBMS 데이터를 하둡 기반 HDFS로 일원화하고 맵리듀스를 이용하여 데이터를 재구성하여 사용자에

게 대량의 데이터를 빠르고 안전하게 추출하는 모델로 하둡 기반 빅 데이터 처리 기법을 적용하였다. [그림 4]는 하둡을 이용하여 데이터 처리를 진행하는 제안시스템의 서비스 구성도로 빅 데이터 플랫폼을 이용한 데이터 처리 흐름 과정이다. [그림 4]의 데이터 처리 흐름을 설명하면 다음과 같다.

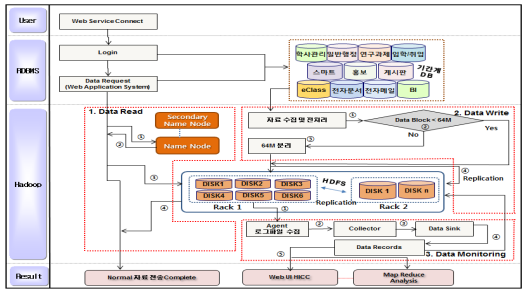


[Fig. 4] Configuration of the Proposed system Service Using Big Data Framework

첫째, 기관 내 정보시스템 데이터는 RDBMS에서 일정 시간을 기준으로 자동화 스케줄러에 의해 데이터 전처리기로 전송한다. 데이터 전처리기에 수집 된 자료는 맵리듀스(MapReduce)를 이용하여 하둡(Hadoop) HDFS 파일시스템으로 일원화된다. 일원화된 하둡 HDFS는 하나의 파일시스템으로 데이터를 관리하게 된다. 둘째, 하둡 HDFS는 안정된 데이터 관리를 위해 별도의 분산 DB 구성도 가능하다. 이 때 사용되는 DB는 NoSQL 기반의 DB를 이용하여 하둡 HDFS로 전달되어 저장, 보관한다. 셋째, 사용자는 웹 정보시스템을 통해 서비스를 요청을 한다. 사용자 요청 서비스 이벤트가 발생되면 웹 정보시스템은 하둡 HDFS에 데이터 처리 요청을 전송한다. 하둡 HDFS는 사용자의 정상적인 요청이 확인하고 사용자에게 결과 값을 전송한다. 넷째, 사용자 요청 값은 사용자에게 전달될 때 128비트 암호문인 AES(Advanced Encryption Standard) 암호화 알고리즘을 사용한다.

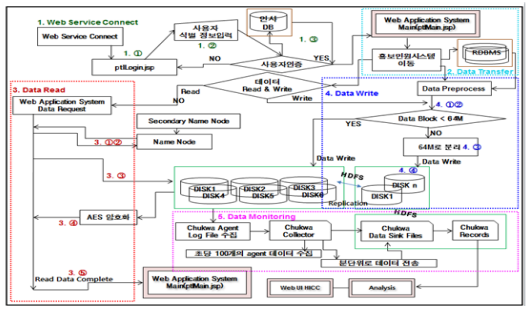
3.3 빅 데이터 처리의 모델 기반의 제안시스템 설계

본 논문의 빅 데이터 처리 기술을 대학 내에서 운영되는 웹 민원시스템과 자산관리시스템의 데이터 저장과 조회 과정에 설계, 적용하였다.

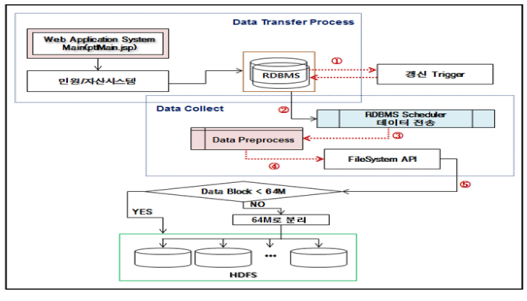


[Fig. 5] Flowchart of the Proposed system Using Big Data Framework#1

[그림 5]는 제안시스템 서비스 흐름도로 본 논문의 전체적인 서비스 흐름을 도식화하였다. 사용자가 데이터를 Request, Write하는 과정 그리고 사용자 서비스를 지원을 위해 RDBMS에서 하둡HDFS 파일시스템으로 데이터를 이관하는 과정을 표현했다. [그림 6]은 본 제안시스템의 서비스 절차 순서도로 사용자가 웹 서비스에 접속하여 원하는 데이터를 읽기(Read), 저장(Write), 상태확인 등의 과정을 순서도로 표현하였다. 전체적인 흐름을 설명하면 다음과 같다. 1. Data 전송 2. Data 저장 3. Data 조회 4. Data 암호화 및 상태확인 등 4가지 서비스 흐름으로 설명 된다.



[Fig. 6] Flowchart of the Proposed system Using Big Data Framework#2

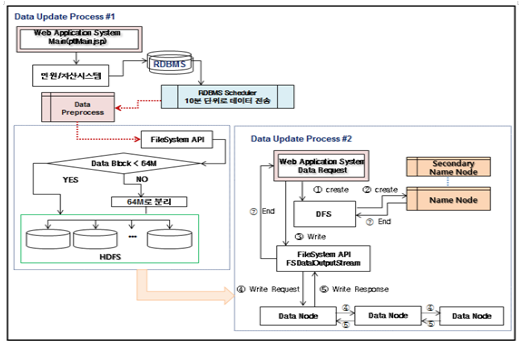


[Fig. 7] Data Transfer Process

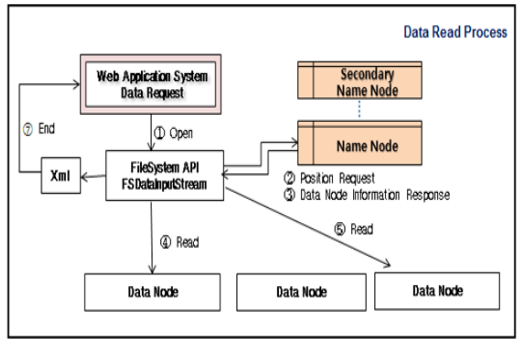
1. Data 전송은 RDBMS에서 데이터 전처리기(Preprocessing)로 데이터를 전송하는 과정이다. RDBMS에서 데이터 갱신 트리거가 발생되면 RDBMS 스케줄러(Scheduler)에 의해 갱신된 RDBMS의 자료를 하둡 HDFS 파일시스템으로 데이터를 전송한다. 데이터 중계기인 데이터 전처리기는 변경된 RDBMS 데이터가 전송되면 즉시 하둡 HDFS 파일시스템으로 데이터를 재전송한다. [그림 7]은 Data 전송 과정을 설명한 흐름도이며 진행은 다음과 같다. 첫 번째 과정으로 RDBMS에서 데이터 갱신 트리거가 발생되면 RDBMS 스케줄러(Scheduler)에 의해 갱신된 RDBMS의 자료를 하둡 파일시스템으로 데이터를 전송한다. RDBMS Scheduler는 10분 내에 RDBMS 데이터를 xml로 변환하여 Data 전처리 서버로 전송한다. 두 번째 과정으로 xml 파일로 수집된 데이터는 데이터 중계기 역할을 하는 데이터 전처리기에서 하둡 HDFS Stream Write API를 이용하여 데이터를 64MB 고정길이 블록으로 만들어져 하둡 HDFS에 저장한다. 또한, 하둡 HDFS에 의해 저장되는 데이터는 기본적으로 3개가 복제되어 인근노드 또는 임의의 노드에 분산 저장된다. 본 수행과정은 [그림 6]의 2. 데이터 전송 과정으로 설명된다.

2. Data 저장은 사용자에 의해 해당 정보서비스 모듈의 데이터 변경 작업이 진행되는 과정이다. 사용자가 사용자 아이디와 암호를 이용하여 웹 시스템에 접속하면 인증검사와 유효성검사 진행 후 사용자는 해당 서비스 모듈로 이동하여 변경데이터를 발생시킨다. 변경데이터는 먼저 RDBMS에 저장되며 RDBMS에서는 RDBMS 스케줄러에 의해 10분 간격으로 변경 된 데이터가 xml로 변환되어 Data 전처리 서버로 전송된다. 스케줄러에 의해 xml 파일로 변환 된 데이터는 하둡(Hadoop) HDFS Stream Write API를 이용하여 데이터를 64MB 고정길이 블록으로 만들어져 하둡 HDFS에 분산 저장한다. 이 때 전송되어 저장되는 데이터의 파일 사이즈는 64MB 단위로 나누어져 3개가 복제되어 인근노드 또는 임의의 노드에 분산 저장된다. Data 갱신(Update) 과정도 이와 동일하게 처리된다. [그림 8]은 데이터 저장(Write) 및 갱신(Update) 과정을 설명한 흐름도이다.

3. Data 조회는 사용자에 의해 해당 정보서비스 모듈의 데이터 읽어오는 과정이다. Data 저장과 동일하게 사용자는 사용자 아이디와 암호를 이용하여 웹 정보시스템에 접속하고 사용자 인증 검사와 암호 유효성 검사를 진행한다. 사용자 인증이 완료되면 사용자는 본인이 원하는 서비스 모듈로 이동하여 해당 정보서비스 모듈의 정보 조회를 요청한다. 데이터 정보 조회를 요청받은 웹 정보시스템은 기존에는 RDBMS가 아닌 하둡 파일시스템의 네임 노드에게 사용자 요청 정보를 전달한다. 하둡 네임노드는 웹 시스템에 사용자 요청 정보를 전달하고 이때 전달되는 정보는 실 데이터가 저장되어 있는 데이터 노드의 정보를 전달한다. 데이터 노드의 정보를 전달받은 웹 시스템은 실 데이터 노드를 검색하여 데이터 암·복호화를 거친 후 사용자에게 안전한 정보서비스를 제공한다.



[Fig. 8] Data Write & Update Process



[Fig. 9] Data Read Process

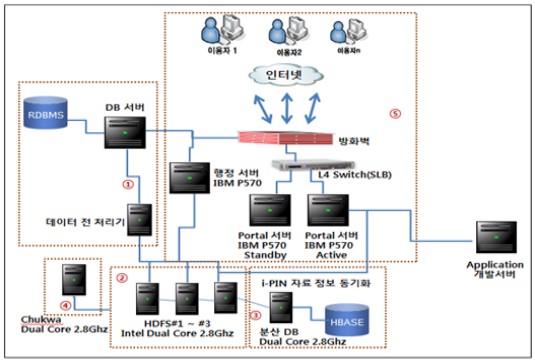
[그림 9]는 Data 조회 과정을 설명한 흐름도이며[그림 8,9]는 [그림 6]의 4. Data 저장과 3. Data 조회 과정을 상세 도식화하였다.

4. Data 암호화는 사용자에게 요청된 Data 조회 정보를 3.의 Data 읽기 과정을 거쳐 사용자에게 결과 값을 전

달하며 결과 값을 암호화하여 사용자에게 전달하는 과정이다. Data 상태 확인은 네트워크상에 분산되어 있는 각 데이터 노드의 로그정보, 작업 진행 상황, 프로그램 상태 정보, 사용자 사용내역 등을 기록하고 상태정보 모니터링에 이용된다.

3.4 빅 데이터 처리의 모델 기반의 제안시스템 구현

본 논문에서는 다형화 되는 데이터 증가로 인한 기관 내 RDBMS 간의 데이터 처리에 관심을 두고 사용자에게는 신속한 정보서비스와 효율적인 데이터 관리를 위해 기존 RDBMS 데이터 운영 방식에서 벗어나 확장성 있는 데이터 운영에 관심을 두었다. 더불어 경제적 비용을 최소화하며 데이터 증가로 인한 서비스 지연 현상을 개선하여 빠른 데이터 처리에 초점을 두고 본 제안시스템을 구현하였다. 다음의 [그림 10]은 본 논문의 빅 데이터 프레임워크 기술을 이용한 데이터 처리 구현 과정을 한눈에 보이도록 서비스 그룹별로 도식화한 시스템 구성도이다. [그림 10]을 간단히 설명하면 ①데이터 수집 ②데이터 조회 및 처리 ③데이터 백업 ④데이터 모니터링 ⑤데이터 암호화 과정으로 모듈별 처리 과정으로 구분하였다.



[Fig. 10] Conceptual Diagram for each Service Module Using Big Data Framework

3.4.1 웹 서비스 접속

사용자는 웹 서비스에 접속한다. 이는 허용된 사용자임을 확인하는 과정으로 Application은 인터넷 웹 서비스에서 입력받은 사용자 정보로 로그인을 요청한다. 다음의 [그림 11]는 웹 서비스 로그인을 구현한 화면이다. [그림 12]은 사용자 접속 승인이 완료되면 사용자에게 서비스되는 메인 화면이다.

3.4.2 데이터 조회

데이터 조회는 RDBM에서 하둡 HDFS로 파일을 이관 후 데이터를 읽어오는 과정이다. 정보 조회를 요청받은 웹 정보시스템은 하둡 HDFS 파일시스템의 네임 노드에게 사용자 요청 메타 정보를 확인하여 실 데이터가 저장되어 있는 데이터 노드의 정보를 읽어온다. 데이터 노드의 정보를 전달받은 웹 시스템은 실 데이터 노드를 검색하여 사용자에게 서비스를 제공한다. [그림 13]는 민원정보시스템에서 읽은 텍스트와 첨부파일의 실사용 예이다. [그림 14]는 이미지 파일을 읽어 실제 구현시스템에서 조회되는 웹 정보시스템 화면이다.

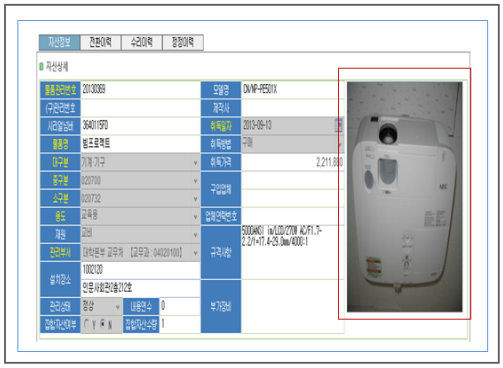
[Fig. 11] Web Service Access Initial Screen

[Fig. 12] Main Screen of Civil Service Using Big Data Framework

3.4.3 데이터 저장

Data 저장은 RDBMS 데이터를 데이터 전처리기로 전송한 후 다시 하둡 파일시스템 HDFS에 분산 저장되는 과정이다. 사용자에 의해 데이터 추가 또는 갱신 데이터가 발생되면 데이터는 먼저 RDBMS 정보시스템에 먼저 저장된다. RDBMS 스케줄러에 의해 데이터 변동 이벤트를 자동으로 감지하면 RDBMS Scheduler에 의해 일정 과정을 거친 후 데이터를 하둡 HDFS로 저장한다.

[Fig. 13] Reading Text and Attached File from Civil Service System



[Fig 14] Reading Image Data from Asset system

4. 결 론

빅 데이터의 활용은 이제 IT 전략의 최우선 과제로 중요자리를 차지하고 있다. 또한, 빅 데이터는 단순히 데이터 용량만을 의미하는 것이 아니라 새로운 기술력의 잠재적 필수 요구 사항으로 자리를 잡고 있다.

본 논문의 데이터 처리 모델은 빅 데이터 기술을 응용한 데이터 관리 모델이다. 기존 RDBMS로 운영하던 데이터 관리를 빅 데이터 처리 기술인 하둡을 이용하여 확장성 있는 데이터 운영과 신속한 웹 정보서비스 지원에 중점을 두고 기존 RDBMS 환경의 데이터를 추출하여 하둡 데이터 관리로 효율적 데이터 처리 과정을 설계하였다. 현재까지 사용하는 RDBMS 환경의 데이터를 빅 데이터 처리 기본 기술인 하둡 클러스터 HDFS 파일시스템으로 일원화하고 각각의 정보시스템에서 수집된 데이터를 사용자에게 안전하게 전달하는 웹 정보서비스 모델을 개발하였다. 과거의 데이터 처리는 정해진 정형화 된 데이터를 처리했다면 현재는 사진, 동영상, 음악, 지도 등과 같이 다양한 데이터를 통합적으로 다루고 있다. 이런 점에서 볼 때 본 제안시스템은 불특정인 다수가 접속하여 수시로 데이터를 읽거나 첨부파일을 다운로드하는 웹 사이트의 비정형 데이터의 조회에서 효율적인 사용예가 되어 주었다. 무엇보다 본 제안시스템은 대용량 파일 저장이 가능한 분산 파일시스템을 사용하므로 클러스터 구성이 가능하고 멀티노드의 부하를 분산 처리하므로 시스템의 과부하나 병목현상을 줄여주는 장점이 있다. 또 하둡은 오픈소스로 제공되므로 경제적 비용을 줄여주는 큰 장점이 있다. 예컨대 하둡과 RDBMS 비용을 분석하면 운영비용이 약 3배 이상의 차이가 발생된다. 하둡은 테라바이트 약 4,000달러 정도인 반면에 RDBMS는 약1만 4천 달러의 비용이 소요되기 때문에 우리는 향 후 하둡에 주목하지 않을 수가 없을 것이다[17].

결과적으로 본 논문에서 사용된 하둡을 이용한 데이터 처리 기법은 기존 RDBMS 운영 환경을 개선하여 효율적인 데이터 운영과 신속한 데이터 처리가 가능하게 했다. 아울러 기존 RDBMS의 재구성 없이 서버나 데이터베이스 등 인프라 시설에도 추가적인 비용이 소요되지 않는 효율적인 데이터 운영이 가능했다. 이는 현재와 같은 다양한 컴퓨팅 환경의 각종 RDBMS 정보시스템을 보유하고 있는 기관에게 경제적 비용의 감소와 신속한 정보서비스를 제공하게 될 것이다.

References

[1] Y. J. Song, *"Policy Challenges for the Future of Data-Based Country Strategy"*, NIA , IT Future Strategy No. 3, Apr 2013.

[2] G. S. Hang, *"Big Data Platform Strategy: Big Data is Changing Business Platform Future Revolution"*, Electronic Times, (pp. 83-97, 101-105, 193-203), 22013.

[3] M. R. Choi, *"United States · Japan: Big Data R&D Strategies and Country of the Corresponding Problem"*, Nipa, IT R&D Policy Review, Mar 2013.

[4] Ms Park Presidential Election Camp, *"Creative Economy"*, New World, 2012.

[5] DongA, *"[2012 General Election-Big Data Presidential Election Campaign] 'Big Data 'Election Period"*, Available From: http://news.donga.com/3/all/20120202/43722588/1, Feb, 02, 2012.

[6] Seth Gilbert, Nancy Lynch, *"Brewer's Conjecture and the Feasibility of Consistent, Available, Partition-Tolerant Web Services"*, ACM SIGACT, (pp. 51-59), vol 33 Issue 2, (accessed June, 2002).
[7] Anonymous. http://develop.sunshiny.kr/883?category=50, 2013.
[8] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung, *"The Google File System"*, 19th ACM Symposium on Operating Systems Principles, Lake George, NY, [Online] Available: http://research.google.com/archive/gfs.html, (accessed Oct, 2003).
[9] Anonymous. *"Big Data at the Speed of Business"* http://www-01.ibm.com/software/data/bigdata/.
[10] STAMFORD, Conn, http://www.gartner.com/newsroom/id/1731916, June 27, 2011.
[11] McKinsey Global Institute, *"Big Data: The next frontier for innovation, competition. and productivity"*, McKinsey Global Institute, 2011.
[12] PHILIP CARTER, *"Big Data Analytics: Future Architectures, Skills and Roadmaps for the CIO"*, WHITE PAPER, IDC sponsored by SAS, 2011.
[13] Vitaly Friedman, *"Data Visualization & Infographics"*, Graphics, Monday Inspiration, January 14th, Jan, 2008.
[14] Anonymous. "Big Data Era-Hadoop", http://cfic.tistory.com/5, May 25, 2012.
[15] J. Dean, S. Ghemawat *"MapReduce: Simplified Data Processing on Large Clusters"*, Communications of the ACM, vol. 51, No. 1, Jan, 2008.
[16] Colin White, *"MapReduce and Data Scientist"*, BI Research, 2012.
[17] Brian Proffittm, *"Cost Analysis of Hadoop and RDBMS…Grenada is a Three-Fold Difference in Operating Costs."*, IDG KOREA, Technology Trends, Jan, 12, 2012.

김 현 주(Hyun-Joo Kim) **[정회원]**

- 2010년 2월 : 단국대학교 정보통신대학원 정보통신학과 (공학석사)
- 2014년 2월 : 단국대학교 대학원 전자·전기공학과 컴퓨터응용 전공 (공학박사)
- 1999년 3월 ~ 현재 : 협성대학교 전산정보실

<관심분야>
빅 데이터, 정보보안, i-PIN, 디지털포렌식, IT융합